1. A window into NER

(a) i. Two examples -
- The weather in New York is pleasant. [New York = LOC]
  The New York Times Company was founded in 1851. [New York Times Company = ORG]
- Austin Waters is my friend. [Austin Waters = PER]
  Paddleboaters took to Lake Austin waters. [Lake Austin = LOC]

ii. It is important to use features apart from the word itself to predict named entity labels because the context around the word enables us to disambiguate among multiple named entity labels that can apply to that word.

iii.
- Capitalization of the first letter of the word
- Window of k words around the word and their POS tags

(b) i.

$$e^{(t)} \in R^{1\times(2w+1)D}$$
$$W \in R^{(2w+1)D\times H}$$
$$U \in R^{H\times C}$$

ii. For each t,

$$e^{(t)} => O((2w+1)D)$$
$$h^{(t)} => O((2w+1)wDH)$$
$$\hat{y}^{(t)} => O(HC)$$

Computational complexity of predicting labels for a sentence of length T is $O((2w+1)DHT + HCT)$

(c) Coding

(d) Analyze the predictions of your model using the files generated.

i. Best development entity-level F1 score = 0.85

Token-level confusion matrix

gold\guess	PER	ORG	LOC	MISC	O
PER	2926.00	43.00	81.00	16.00	83.00
ORG	109.00	1687.00	115.00	55.00	126.00
LOC	22.00	71.00	1931.00	21.00	49.00
MISC	37.00	49.00	47.00	1030.00	105.00
O	30.00	46.00	21.00	29.00	42633.00

- The model is good at detecting the "Null" class
- ORG is misclassified as O, LOC and PER in decreasing frequency. ORG seems to be the class which the model is most confused about
- LOC is sometimes misclassified as ORG
- MISC is sometimes misclassified as NULL
- The model is pretty good at recognizing PER and O (Null) labels

ii. 2 modeling limitations of the window-based model
- Window size 1 forces the model to make predictions based on a very local context of one word on either side of the current word. This restricts the model's ability to disambiguate the correct word sense.
  Due to window size 1, the word "Ashes" has context "the Ashes is" which does not provide any disambiguating information. It is misclassified as LOC. Similarly, the word "Test" has context "the Test and" which causes it to be classified independent of the context that it is followed by "County..." which might have helped provide more useful information for classification.

```
Australia will defend the Ashes in
y*: LOC       O    O      O   MISC  O
y': LOC       O    O      O   LOC   O
x : a six-test series against England during a four-month tour
y*: O O        O      O       LOC     O      O O          O
y': O O        O      O       LOC     O      O O          O
x : starting on May 13 next year , the Test and County Cricket Board
y*: O        O  O   O  O    O    O O   ORG  ORG ORG    ORG     ORG
y': O        O  O   O  O    O    O O   MISC O   ORG    ORG     ORG
```

- The model cannot consider non-local context in making it's decisions

```
x : Pint is looking for people . Pint Corp is doing well .
y*:
y': PER  O  O       O   O      O ORG  ORG  O  O     O    O
```

2. Recurrent neural nets for NER

(a) i. How many more parameters does the RNN model in comparison to the window-based model?
The additional parameters come from the dependence on the previous time step hidden layer $h^{(t-1)}W_h$ i.e. $W_h \in R^{H \times H}$ is the additional parameter

ii. What is the computational complexity of predicting labels for a sentence of length T (for the RNN model)?
For each time step, $O(D \times H + H \times H + H \times C)$
If $C << H$ and $C << D$, then $O(D \times H + H \times H)$ For t time steps, $O(DHT + H^2T)$ computational complexity for prediction

(b) i. Name at least one scenario in which decreasing the cross-entropy cost would lead to an *decrease* in entity-level F1 scores.
In a multi-word enitity, e.g. New/LOC York/LOC - If our prediction changed from New/MISC York/MISC to New/MISC York/LOC, then the cross entropy error decreases since we predicted one more word correctly. But the change also implies that we are predicting two entities incorrectly (New/MISC and York/LOC) as opposed to one entity previously (New/MISC York/MISC). This decreases the precision, while recall remains the same. Decreased precision decreases F1.

ii. Why it is diffcult to directly optimize for F1?
F1 is non-convex. If F1 is computed at the entity level, computation of F1 at each epoch would require doing prediction over the entire corpus which may be very expensive. And since the entire corpus is required, it is not possible to batch (stochastic) and/or parallelize the training.

(c) Coding

(d) i. How would the loss and gradient updates change if we did not use masking? How does masking solve this problem?
$J = \sum_{t=1}^{M} CE(\hat{y}^t, y^t)$ would be the loss if we did not use maksing. If sentence length $T < M$, then there will be additional predictions for padded tokens and the loss on those padded tokens will be included in the loss computation. Similarly, gradients with respect to W_h will be updated based on errors on the padden tokens as well. Masking makes the padded tokens error to zero. So, neither loss nor gradient updates are affected by the padded tokens prediction error.

ii. Coding

(e) Coding

(f) Development F-1 score = 86 %

(g) i. Describe at least 2 modeling limitations of this RNN model and support these conclusions using examples from your model's output.

ii. For each limitation, suggest some way you could extend the model to overcome the limitation.